Behaviour of Recent Aesthetics Assessment Models with Professional Photography

Mathieu Chambe, Rémi Cozot, Olivier Le Meur

Behaviour of Recent Aesthetics Assessment Models with Professional Photography

Mathieu Chambe, Rémi Cozot, Olivier Le Meur

Abstract—Aesthetic quality assessment for photographs is an important research topic since it can be used by a number of applications, such as image database management or image browsing. In 2012, the Aesthetic Visual Analysis (AVA) dataset has been proposed. Those 255,000 aesthetically annotated images are a key ingredient for training and for testing new models for aesthetics prediction. As AVA dataset is mainly composed of competitive photographs, we evaluate whether or not those computational models of aesthetics generalize well and perform well over professional photographs. We notice that the different models we test behave quite differently. Besides, we fine-tune the model using professional photographs and the results show that this process is effective.

Keywords—Aesthetics Assessment; Deep Learning; Overfitting; Professional Photography.

I. Introduction

Assessing the aesthetic quality of images using computational models has been a problem in the computer vision field for many years. This has some applications in image sorting for databases management, or in aesthetics-driven image processing for example. Yet, automatically scoring the beauty of an image is still a difficult problem. Compared to the problem of image quality assessment [1], [2], aesthetics assessment has to be measured by using high level features, which are hardly described by common low level features. Beyond this point, the problem is even trickier since the aesthetic quality of an image is a highly subjective value.

For predicting the aesthetic quality of an image, many computational models have been proposed. The first models are based on specific photographic rules, such as the rule of thirds related to image composition, or narrow depth of field [3]. Performances of such models are however rather limited. We are currently witnessing a new breakthrough in this field thanks to the emergence of huge datasets (e.g. AVA [4]) and new machine learning methods relying on deep learning algorithms. Thanks to deep networks, trained over millions of images [5], it is now possible to get a large number of features able to describe complex and abstract patterns. A more thorough study on recent models can be found in [6].

In 2012, Murray et al. [4] proposed AVA, a dataset specifically designed for aesthetics assessment methods. This dataset composed of more than 250,000 aesthetically annotated and scored images from the photography website www.dpchallenge.com greatly helped research in this domain. However, this dataset is mainly composed of competitive photographs aiming to be shown in juried exhibitions, to be published in specialized photography magazines or web sites, and to compete for recognition and prizes, as described in [7]. For the specific case of AVA, the scores range from 1 (i.e. ugly) to 10 (i.e. beautiful). The mean score is 5.10; the maximum and minimum scores are 8.52 and 1.81, respectively. As current models are trained over the AVA dataset, we could say that current aesthetic models are mainly dedicated to scoring competitive photography.

We can consider two different usages of photography: competitive and professional.

Competitive photographs, as mentioned previously, are beautiful images that should be liked by a very large audience [7]. Competitive photography existed before Internet, but it becomes very popular with specialized photo social sharing site such as www.instagram.com, www.DPchallenge.com, www.flickr.com, etc. To be liked by a larger number of people, competitive photography usually follows classic aesthetics rules and is very aesthetically conservative.

Professional photographs aim to be seen by a large number of people and aim to convey a message or an emotion. Since the beginning of professional image creation (including paintings and photographs), an image of high aesthetic quality is in most of the cases more efficient to convey the intent, message or emotion. In this professional case, aesthetics means well designed technically speaking, but does not necessarily mean that the photography is pleasant. Indeed, the photograph objective can be to shock in order to produce a reaction. Professional photography can further be classified into two main genres: photojournalism (war photography, sport photography) and product photography (fashion photography, real estate or architecture photography, etc.). The former endeavors to capture an instant or its related emotion, while product photographs aim to promote a product to make it desirable.

The main difference between those categories is the aim of the photograph. Competitive photography aims at being pleasant for a majority of people, therefore high aesthetic quality is the end goal of such photography. On the other hand, professional photography aims at conveying the intent of the photographer or its commissioner. Complying with common aesthetic rules in this case is only one mean among others to achieve such goal.

In this paper, the objective is to test whether or not

All authors are affiliated with Univ Rennes, CNRS, IRISA
Mathieu.Chambe@irisa.fr
Remi.Cozot@univ-littoral.fr
Olivier.Le_Meur@irisa.fr

aesthetics prediction models perform and generalize well on different categories of professional photography. This would allow us to quantify the coverage of those models – kinds of photographs accurately rated by the model –, and thus would help us improve the coverage, to achieve a better accuracy in prediction. For this purpose, we put to the test the recent model NIMA [8] as well as the ranking network model [9]. After assessing the general behaviour of the models on some professional photographs, we fine-tune the model with one of the datasets to increase the coverage of the model. This process effectively improve the model.

The rest of this paper is structured as follow: related works are presented in section II; section III presents the models and the different professional datasets we use in the experiments; section IV presents the experiments themselves; section V presents the results of the experiment and finally, section VI is a conclusion that sums up the findings and proposes some future work.

II. Related Work

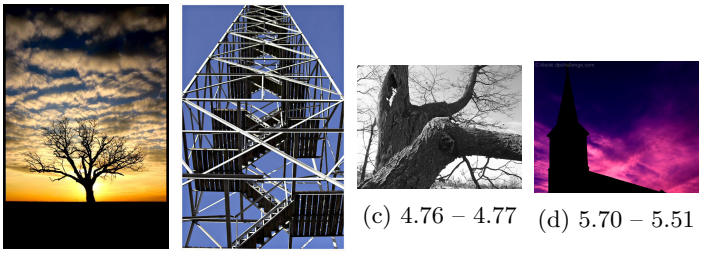
Very few studies on the relevance of training datasets for aesthetics assessment have already been done.

Carballal et al. [10] recently exposes the limits of existing training datasets, and creates their own dataset composed of images from www.dpchallenge.com. Those images are rated in three ways: the mean score given to the image by users of www.dpchallenge.com; an aesthetic score given by observers in a controlled environment; and a preference score given by observers in a controlled environment. This is the first and only dataset for aesthetics prediction having scores from several populations. However, this dataset was not compared to the most used dataset today (AVA) and contains only 1,000 images.

Some models were proposed to counter the bias caused by the imbalance of AVA [4]. In that optic, Jin et al. [11] propose the weighted CNN architecture. This is an architecture based on the VGG-16 network, but differs from the VGG-16 network with the cost function. Instead of using a classic mean square error function, they add weights corresponding to the frequency of apparition of the score of the image. To compute those weights, the authors use the histogram of notes of the training dataset (in this case, AVA). The weight $w(I)$ of an image I is proportional to the inverse of the number of images having the same average score as I. The cost function is then defined as

$$C(\hat{\mathbf{y}}) = \sum_{I \in \mathcal{I}} w(I)(\mathbf{y_0}(I) - \hat{\mathbf{y}}(I))^2 \qquad (1)$$

with $\mathcal{I}$ the set of training images, $\hat{\mathbf{y}}$ the vector composed of the computed scores of each image in the dataset and $\mathbf{y}_0$ the vector composed of the ground truth score of the images. This method allows to have greater weights on images which have low presence on the score scale. This leads to a network having a greater range of action and being able to note accurately images with very high ($s > 6$) or very low ($s < 4$) scores.



(a) 5.81 – 6.24 (b) 5.34 – 5.24 (c) 4.76 – 4.77 (d) 5.70 – 5.51

Fig. 1: Different images with scores from the third-party implementation – from the original model NIMA [8].

III. Testing models and datasets

The proposed experiment relies on testing two recent computational aesthetics model, namely NIMA [8] and the ranking network proposed by Kong [9] with datasets of professional photography. We present the models and the images in this section, as well as the fine-tuning process.

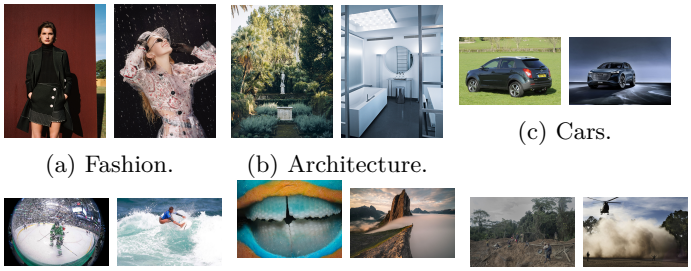
A. NIMA model

In the following, we present the NIMA architecture we use and the training process. More details can be found in the original paper. The main feature of NIMA model is its ability to rely on existing pre-trained deep networks. It consists in replacing the last layers of an image classification network with a fully-connected layer, and then train only the final layer. The model is first completely trained on ImageNet, and then fine-tuned using a specific dataset. Two types of NIMA model are also proposed: one for predicting the aesthetics score of an input image and a second one for predicting the quality of the input image. The former is trained over AVA dataset whereas the latter is trained over TID2013 dataset. In this study, we focus only on the aesthetics prediction models.

NIMA model relies on existing architectures, two of which are Inception [12] and MobileNet [13]. Those are architectures for image classification. NIMA adapts those models to the problem of aesthetics assessment. According to the authors, the model based on Inception is more accurate, but slower than the model based on MobileNet.

A third-party implementation is available online [14]. This implementation also proposes a NIMA model based on NasNet [15], another neural network for image classification. NasNet-based NIMA model performs better on AVA than the models based on Inception or MobileNet. NIMA outputs a distribution of notes so the performance of this model is measured with an Earth Mover Distance (EMD). NasNet gets 0.067 EMD while Inception gets 0.070 EMD and MobileNet gets 0.080 EMD. In the following, we perform our study on the most relevant architectures which are Inception and NasNet.

Before going further, we have checked that the behaviour of the third-party model and the NIMA model are similar, although the scores of images are not exactly the same. Figure 1 illustrates some examples with both predicted scores. On a sample of 4662 images from AVA, we



(a) Fashion. (b) Architecture. (c) Cars.

(d) Sport. (e) National Geographic. (f) War.

Fig. 2: Different sample images from professional datasets. We show here the worst (on left) and the best (on right) of each category according to NIMA.

achieve a correlation coefficient of 0.581 between ground truth and our implementation, which is close to the 0.636 announced by the original article.

B. Ranking network

The ranking network was proposed by Kong et al. in 2016 [9]. The input of the model is two images passing through two identical networks. The output is a score for each image (on a scale from 0 to 1) and a ranking between the two images. Besides, the model outputs several characteristics of the images (compliance with the rule of thirds, presence of symmetry, of vivid colors, etc...) that were used to compute the final score. As the model needs this information for training, the authors also devised a new training database called AADB. The implementation was provided by the authors themselves on their GitHub [16]. More details can be found in the original paper.

C. Datasets of professional photography

In this section, we present the photograph datasets we use to test the models on professional photography. In order to cover a wide range of professional photograph, we use 6 datasets corresponding to different photography genres. Figure 2 illustrates a sample of images of these photography genres:

Fashion category contains photographs coming from various editorial fashion photo-shoots and published in fashion magazines during the year 2018. These photos are captured by professional photographers in collaboration with magazine art director. The photographs are of different aesthetic styles (black and white, color, high key, low key, etc.) These images are not only of high aesthetic quality but also have an artistic dimension. We collect 1373 images.

Architecture category contains real estate, indoor design and architecture photographs published in *Architectural Digest Magazine*. These photographs have been captured by professional photographs and promote the beauty of architecture. They are of high aesthetics quality but in a more classic manner than fashion ones. We collect 117 images.

Cars category contains photographs done by various car manufacturers to advertise their new cars. Due to marketing strategy, the aim of these photographs is to promote different values such as power, robustness and sometimes beauty. We collect 109 images.

Sport category contains photographs of various sports from the French journal *L'équipe* dedicated to sport news. These images have been captured by professional photographers and try to capture crucial moments. We collect 155 images.

War category contains war photographs from the photography agency called *Agence VU*. Obviously the first aim of these photos is not to make aesthetically pleasant photos, but to tell the truth about war. We collect 138 images.

National Geographic category contains wildlife and landscape photographs from the National Geographic website. Similarly to the previous datasets, the photographs have been captured by professional photographers. The usual objective of these photographs is to show the beauty of Earth and wild life. We collect 110 images.

D. Overview of models

Figure 3 shows histograms of predicted scores for images from the datasets presented in section III-C.

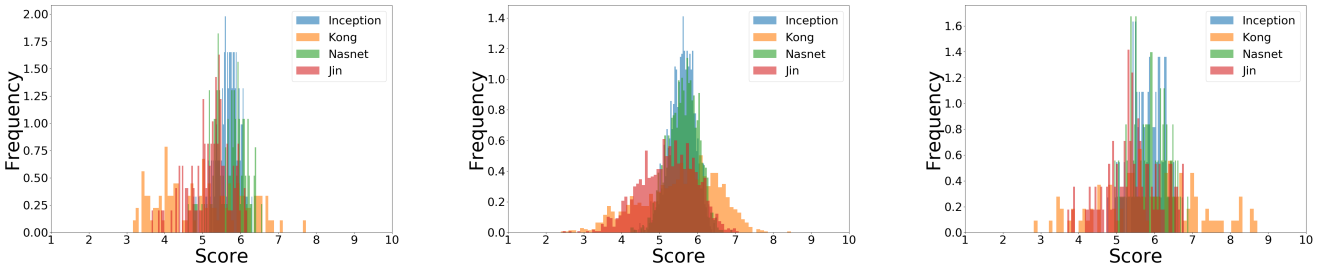
The first observation is that histograms for the two versions of NIMA models are very similar, and present a strong overlap (green/blue on Figure 3). In addition, both histograms can be efficiently represented by a 1D Gaussian distribution with a rather small dispersion ($\sigma = 0.37$ in average). Regarding the ranking network (orange on Figure 3), the distribution is more spread and does not fit well a Gaussian distribution. We then observe a strong discrepancy between the distributions of the two models. This is especially noticeable for War and Sport categories. The War histogram (Figure 3 (e)) is well below the average score while the Sport histogram (Figure 3 (f)) is well above the average. The scores from War were expected: as we said in section III-C, their main goal is not to be appealing. Therefore, they tend to have lower aesthetic values. On the other hand, the sport images are quite colourful and with a low depth of field. These are two qualities that the ranking network assesses. Therefore, they get higher scores than average. This difference between both histograms proves that some kind of over-fitting is present, especially with the NIMA models.

IV. Presentation of experiments

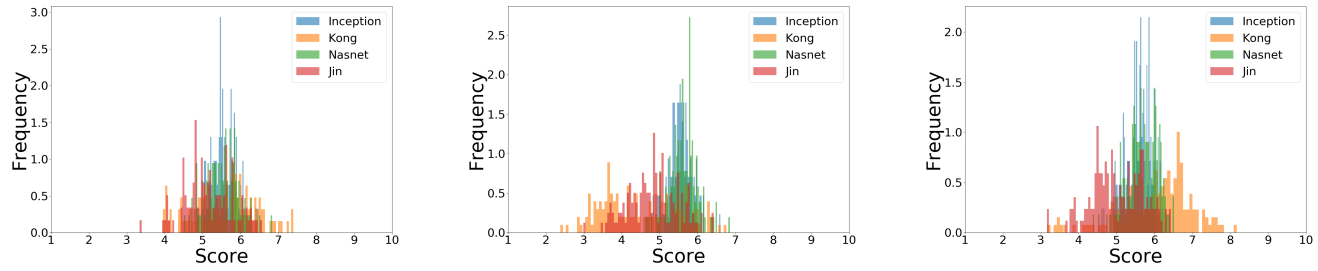
A. Hypotheses

We conduct our experiments under some assumptions. These assumptions are presented and motivated in this section.

The images from the category Fashion are expected to have high aesthetic quality, and therefore high scores. This assumption is reasonable as not only were these photographs taken by professional photographer, they were also published, which proves that they are acknowledged to



(a) Architecture. (b) Fashion. (c) National Geographic.



(d) Cars. (e) War. (f) Sport.

Fig. 3: Histograms of scores for the four models and for different datasets.

be efficient. The aim of such images is to promote the value of fashion products, so we can argue that the end goal specific to the Fashion category is to have high aesthetic quality.

On the other hand, the images from the category War are expected to have low aesthetic value (according to common aesthetic sense), and therefore would get rather low scores.

B. Fine-tuning of Nasnet-based model

To address the over-fitting problem, we fine-tune the Nasnet-based model using one of the professional datasets. We consider Fashion, as it is the one containing the most images. We construct a training dataset using the 1373 images from Fashion and 1373 random images of AVA. Among these 2746 images, 300 were set aside and used as a validation set (150 from Fashion and 150 from AVA). As we do not have any ground truth scores for Fashion, we create scores by changing the overall mean score. We make the assumption that, as professional photographs, the images from Fashion must have a high mean score. We devise a method to score images using the score from NIMA as a basis. If the current mean score for the dataset on Nasnet-based NIMA is μ, the ground truth score for each image of score s is given by $\bar{s} = s - \mu + \bar{\mu}$ where $\bar{\mu} = \mu_{high} + 1$ is the new mean score. $\mu_{high} = 6$ is the score corresponding to high aesthetics value according to many previous work [17], [4]. Using this method, we ensure that the fashion dataset has a mean score of $\bar{\mu}$, and therefore, a majority of images from Fashion have a score higher than μ_{high}.

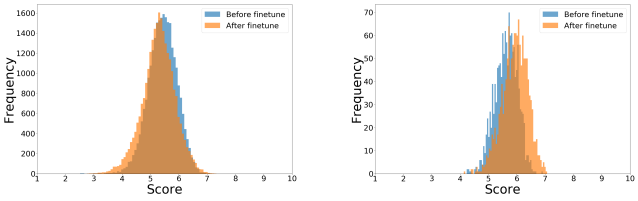


Fig. 4: Influence of the fine-tuning process using the Fashion dataset on Nasnet-based NIMA scores. Results are reported for AVA (left) and Fashion (right).

V. Results

A. Statistical analysis

Table I presents the mean scores and whether the paired t-test for all pairs of scores computed by the tested models are significant or not.

NIMA models.: Results indicate that the mean score of ground truth scores (5.1) is significantly lower than the mean score of tested professional photographs (ranging from 5.52 to 5.84). This difference, although very small on a scale of 10 grades, was expected and statistically significant.

Ranking network model.: As the distributions are not normal, we do not use a t-test, but a Wilcoxon rank-sum test. We observe that the range of scores is greater than for NIMA; the dynamic of scores represents 54.4% of the whole scoring scale, whereas NIMA scores represent only 22.3% of the scoring scale in average. This observation would suggest that the ranking model is more selective for the professional categories.

	AVA	NG	C	F	A	S	W
NIMA NasNet Mean Score	5.10	5.84	5.52	5.63	5.69	5.67	5.62
AVA		***	***	***	***	***	***
Nat.Geo.			***	***	*	**	***
Cars				*	**	*	ns
Fashion					ns	ns	ns
Archi.						ns	ns
Sport							ns
War							
NIMA Inception Mean Score	5.10	5.84	5.55	5.59	5.66	5.57	5.53
AVA		***	***	***	***	***	***
Nat.Geo.			***	***	**	***	***
Cars				ns	**	ns	ns
Fashion					**	ns	ns
Archi.						*	***
Sport							ns
War							
Ranking Net Mean Score	0.455	0.545	0.510	0.503	0.443	0.565	0.386
AVA		***	***	***	ns	***	***
Nat.Geo.			*	**	***	ns	***
Cars				ns	***	***	***
Fashion					***	***	***
Archi.						***	***
Sport							***
War							

TABLE I: Table of paired t-test (or Wilcoxon rank-sum) p-values for different datasets (AVA=Ground truth scores; NG=National Geography; C=Cars; F=Fashion; A=Architecture; S=Sport; W=War) on the three models. The stars are attributed using the p-values: * for $0.05 \geq p > 0.005$, ** for $0.005 \geq p > 0.0005$, *** for $0.0005 \geq p$; *ns* stands for non significant.

B. Does fine-tuning Nasnet-based NIMA model improve the overall prediction capabilities?

Figure 4 presents predicted scores when the Nasnet-based NIMA model is fine-tuned following the procedure described in subsection IV-B. As expected, we notice that the scores of Fashion have increased while AVA scores slightly decreased. The fine-tuned model is then able to better discriminate Fashion photography from competitive photography. It may suggest that models trained over AVA are specialized for competitive photography. However, a simple fine-tuning process allows us to make the model more generic and more relevant.

C. Comparison with weighted CNN

As the proposed fine-tuning process and the weighted CNN [11] have the same goal (reducing the bias caused by the imbalance of AVA), we compare the performance of both networks. The model and weights of Jin et al. is available on their website [18]. We can then compare our Nasnet-based NIMA model fine-tuned with Fashion and the original weighted CNN model. Figure 4 presents our datasets outputs on Nasnet-based NIMA fine-tuned (in orange) and Figure 3 presents the results for the original weighted CNN model (in red).

The histograms from Nasnet-based NIMA (Figure 4 in orange) have smaller dispersion than the ones from the weighted CNN (Figure 3 in red). Therefore, the model from Jin et al. seems to be more able to reduce disparity in high and low score value zones. However, we notice a significant difference in mean for the Fashion category in both models, Jin et al. being the lowest. It is thus possible that the weighted CNN significantly improve the dispersion of the score histogram, but is overall less accurate on the score themselves. We can verify this using a correlation metric.

We compute the mean-square error (MSE) and the Pearson correlation metric with AVA for different models: our Nasnet-based NIMA fine-tuned on the Fashion dataset, the original Nasnet-based NIMA and the weighted CNN proposed by Jin et al. These values are reported on Table II.

	MSE (↓)	Pearson ρ (↑)
Original NIMA	0.387	0.618
Fine-tuned NIMA	0.391	0.596
Jin et al.[11]	0.500	0.585

TABLE II: Mean square error and Pearson correlation metric for different models.

First, we notice that the best model according to both metrics is the original Nasnet-based NIMA. Our fine-tuning process slightly degrades the performances of NIMA on AVA. However, as explained previously, it also significantly improves the performances on the Fashion dataset. The fine-tuning process is thus quite effective and allows for better results on Fashion, and good results on AVA. In terms of correlation, the weighted CNN is worse than NIMA and the fine-tuned NIMA. This shows that our method is more faithful to the ground truth scores. All of this proves that our fine-tuning process is a real improvement over the weighted loss function.

D. Results discussion

Using the fine-tuning method, we manage to increase the score of the Fashion database without modifying too much the scores from other categories. If we assume that the Fashion dataset is mainly composed of high aesthetic quality images, we improve the model accuracy and coverage. However, we can discuss the relevance of our assumption.

The score threshold used as high aesthetic quality is used in previous work as a distinction between professional and amateur photographs. The professional photographs we used in our experiments were chosen because of their relevance. Indeed, not only were these photographs taken by professional photographer, they were also published, which proves that they are acknowledged to be efficient. This shows that our assumptions on the scores (of images from War and Fashion) are reasonable.

VI. Conclusion

In this paper, we present a study based on the models NIMA and the ranking network. We aim to understand how these models behave with other kinds of photography than their training dataset. As the dataset AVA is composed of competitive photographs, we have chosen

six datasets of professional photographs in order to test whether or not the models generalize well.

We observe that NIMA and the ranking network have different behaviours. NIMA gives scores with rather small deviation around the mean, whereas ranking network scores are much more spread on the rating scale. We also notice that, for NIMA, there is a strong discrepancy between the scores of AVA and professional photographs. This is alleviated by the fine-tuning process, but raises other issues, such as the correct way to fine-tune, and perhaps train, the networks.

These observations reflect how far we are from accurately predicting the aesthetics of an image. However, we demonstrate that fine-tuning existing models with professional photography can alleviate the over-specialization of existing models to competitive photography. The next step is to define and to provide to the community a new annotated image dataset of professional photographs.

References

[1] Z. Wang, "Applications of objective image quality assessment methods [applications corner]," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 137–142, 2011.

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.

[3] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, Sep. 2011.

[4] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012.

[14] "https://github.com/titu1994/neural-image-assessment," [Online].

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] Y. Deng, C. C. Loy, and X.Tang, "Image aesthetic assessment: an experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[7] A. Tifentale and L. Manovich, "Competitive photography and the presentation of the self," in *Exploring the Selfie*. Springer, 2018, pp. 167–187.

[8] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[9] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 662–679.

[10] A. Carballal, C. Fernandez-Lozano, N. Rodriguez-Fernandez, L. Castro, and A. Santos, "Avoiding the inherent limitations in datasets used for measuring aesthetics when using a machine learning approach," *Complexity*, vol. 2019, p. 12, January 2019.

[11] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Image aesthetic predictors based on weighted cnns," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2291–2295.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[15] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[16] "https://github.com/aimerykong/deepImageAestheticsAnalysis," [Online].

[17] W. Wang, S. Yang, W. Zhang, and J. Zhang, "Neural aesthetics image reviewer," *arXiv preprint arXiv:1802.10240*, 2018.

[18] "https://ivrlwww.epfl.ch/bjin/project_aesthetics/Image_Aesthetics.html," [Online].